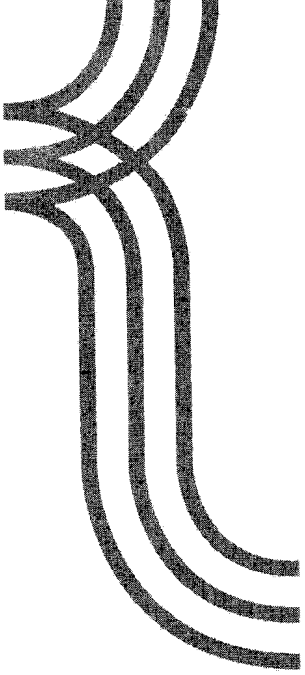
数学之美

吴军 著

JUST PUB

人民邮电出版社
POSTS & TELECOM PRESS



JUST PUB

数学之美

Beauty of Mathematics

吴军 著

人民邮电出版社
北京

图书在版编目（CIP）数据

数学之美 / 吴军著. -- 北京 : 人民邮电出版社,
2012.6(2012.6重印)
ISBN 978-7-115-28282-8

Ⅰ. ①数… Ⅱ. ①吴… Ⅲ. ①电子计算机－数学基础
Ⅳ. ①TP301.6

中国版本图书馆CIP数据核字(2012)第088566号

内 容 提 要

几年前，“数学之美”系列文章原刊载于谷歌黑板报，获得上百万次点击，得到读者高度评价。读者说，读了“数学之美”，才发现大学时学的数学知识，比如马尔可夫链、矩阵计算，甚至余弦函数原来都如此亲切，并且栩栩如生，才发现自然语言和信息处理这么有趣。

今年，作者吴军博士几乎把所有文章都重写了一遍，为的是把高深的数学原理讲得更加通俗易懂，让非专业读者也能领略数学的魅力。读者通过具体的例子学到的是思考问题的方式——如何化繁为简，如何用数学去解决工程问题，如何跳出固有思维不断去思考创新。

数 学 之 美

◆ 著　　吴 军
责任编辑　俞 彬
审稿编辑　李琳骁
策划编辑　周 筠

◆ 人民邮电出版社出版发行　　北京市崇文区夕照寺街14号
邮编　100061　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京铭成印刷有限公司印刷

◆ 开本：720×960　1/16
印张：19　　彩插：1
字数：248千字　　2012年6月第1版
印数：40 001－80 000册　　2012年6月北京第4次印刷

ISBN 978-7-115-28282-8

定价：45.00元

读者服务热线：(010)67132692　印装质量热线：(010)67129223
反盗版热线：(010)67171154

本书谨献给我的家人。

愿科学之精神在国民中得到普及，愿中国年轻的一代涌现更多的杰出专业人才。

出版说明

“数学之美”最初是从2006年起在Google中国的官方博客——谷歌黑板报上连载的系列博客。当时我写这个系列的原因完全是应原黑板报版主吴丹丹女士（现任职于苹果公司）之请，希望介绍一点Google的技术，盛情难却，便勉为其难接下了这个任务。这个任务的难处在于既要介绍Google的技术，又不能泄密。于是我只好采用了仅介绍基本原理尤其是数学原理的方法来写文章。加上我自己对数学比较感兴趣，博士论文也是以数学为主的题目，因此，便写成了介绍我所从事的信息处理领域的数学基础的系列文章。当初我并没有计划写多少篇，只打算有空就抽时间写一点，写到哪儿算哪儿，没时间写就算了。不想刊登了几篇后，受到IT行业广大从业人员的关注和喜爱。在互联网上被转载了上万次，读者有上百万。大家都鼓励我写下去，于是便陆陆续续写了20多篇。

后来我在Google的工作越来越多，同时在公司外还有很多要履行的义务，便很少有精力再写这个系列了。2009年，李开复离开Google，谷歌黑板报的第二任版主崔瑾女士也随他去了创新工场（后来转到豌豆荚）。Google再也没有人敦促我继续为谷歌黑板报写博客了。令我感动的是，这么多年后，还不断有读者关注这个系列，并且时不时地问我是否能把这个系列写完，是否可能出书。因此，从2010年起，我陆续将剩下的几篇写完。

出书比写博客要求高很多。一本好书需要结构系统而文字严谨，为了达到出书的要求，我几乎重写了所有的内容。因此，这本书虽然在每章的标题和主题上与原来的博客相同，但是内容和文字都是新的。希望广大读者，无论是过去读过黑板报上连载系列的老朋友，还是第一次读这本书的新朋友，都能有全新的收获。

在系统性方面，为了便于非IT读者的阅读，我对每个专题都给出了背景介绍；同时，为了起到给从事相关工作的工程师做参考的目的，在一些专题的最后，我都给出了“延伸阅读”一节。非IT读者可以完全跳过这些延伸阅读部分，这样并不会影响阅读其他内容。本书中系统性方面的第二个改进就是调整了章节的位置，以帮助读者阅读。在严谨性方面，我在腾讯工程师王益等人的帮助下，更正了原来博客中的一些错误，并尽可能补充完善很多公式推导的过程。

本书的素材来源于我本人的工作。语言信息处理、互联网技术、数据挖掘和机器学习都是博大精深而又快速发展的领域，我所做的研究工作也只涵盖了其中很小的一部分。对于我没有涉足过的领域，我没有信心也没有资格写。因此，这本书在内容上并没有全面覆盖上述领域，比如对当今数据挖掘领域的算法、互联网上各种推荐系统的数学模型都鲜有提及。对这些内容有兴趣的读者可以查阅相应的书籍文章，也希望今后有这方面的专家能够将自己工作的心得写出来，供大家学习参考。

在《数学之美》成书之际，我要感谢所有那些把我带到数学王国和信息处理领域的人。特别要感谢的是我的父亲，他让我在幼年就对数学和自然科学产生了浓厚的兴趣，并帮助我打好数学基础。接下来要特别感谢的是我在中国和美国的三位导师：王作英教授、库坦普教授和贾里尼克教授，他们三人都是有着很深数学造诣的信息论专家，他们把我领进语音和语言处理的王国，不仅帮助我打下比较深厚的数学功底，而且让我看到了数学的威力和魅力。最后要感谢Google的谢尔盖·布林院士、韦恩·罗森先生和彼得·诺威格博士，他们把我招进Google，让我有机会

进入网络搜索领域。尤其是诺威格博士，作为我在 Google 的直接上级，一直给予我做任何我想做的工作的权力，因此，我才得以精通网络搜索的每一个细节。在 Google，我还要感谢我的同事阿米特·辛格博士，他是我在搜索领域的导师。

在这本书的写作过程中，我得到了很多人的帮助和鼓励。Just Pub 团队的审稿编辑李琳骁先生对书稿进行了数次精心的审读，感谢他的认真和专业。华中科技大学数学与统计学院的周笠教授也拨冗通读了全书，感谢他的鼓励和细心。腾讯公司的王益博士（原 Google 工程师）帮我编辑和校对了书中诸多的公式，并完善了书中的一些细节部分。吴丹丹女士在将我的拙作刊登在谷歌黑板报前，对我的博客进行了润色，尤其是增加了不少画龙点睛之笔。在我将博客编辑成书时，Google 的叶艳女士通读了所有的章节，并给出了修改意见，以保证全书对于非专业读者的可读性。

清华大学的李星教授、李开复博士、本书的出版人周筠女士和原 Google 中国的总监王怀南先生一直是这个博客系列的热心推荐人，并且一直鼓励我将它编纂成书。我以前在清华大学的同事郭进博士是自然语言处理专家，经常和我讨论中文处理的问题，并且给了我很多启发。在此，我向他们表示诚挚的感谢。

同时我要感谢我的家人给我的帮助，特别是我的两个女儿吴梦华和吴梦馨，她俩绘制了全书的许多插图。

最后要感谢所有热心的博客读者以及在互联网上传播这个博客的媒体、网站和个人。希望这本书能够帮助读者从广度和深度上了解信息科学。

吴军

2012 年 4 月于深圳

序言 1

《数学之美》是一本非常值得读的书。这本书展现了吴军博士在他多年的科研经历中对科学问题的深入思考。

我于1991年从美国回到清华大学电子工程系工作，与吴军博士是同事，对他在汉语语音识别方面的深入研究印象非常深刻。后来他到美国工作，出版了一本介绍硅谷的书《浪潮之巅》，使我对他的写作激情和水平有了新的认识。

这些年来我在清华大学教书，一直思考着如何让学生能真正欣赏和热爱科学研究，这将有助于他们深入理解自己所从事的研究的价值，进而能逐渐成长为所在领域的大师和领军人物。在这一过程中，恰好发现了吴军博士在谷歌中国的官方博客 —— 谷歌黑板报上连载的“数学之美”系列文章，我非常欣赏这些文章。因此，在很多场合都建议学生跟踪阅读这个系列的博客文章。今天本书出版，与原先的博客文章相比，其内容的系统性和深度又上升到了一个新的境界。

我读《数学之美》有下面几点体会，与大家分享。

1. 追根溯源

《数学之美》用了大量篇幅介绍各个领域的典故，读来令人兴趣盎然。典故里最核心的是相关历史事件中的人物。我们必须要问：提出巧妙数学思想的人是谁？为什么是“他 / 她”提出了这个思想？其思维方法有何特点？成为一个领域的大师有其偶然性，但更有其必然性。其必然性就是大师们的思维方法。

2. 体会方法

从事科学研究，最重要的是掌握思维方法。在这里，我举两个例子。

牛顿是伟大的物理学家和数学家，他在《自然哲学的数学原理》中叙述了四条法则。其中有“法则 1：除那些真实而已足够说明其现象者外，不必去寻找自然界事物的其他原因”。这条法则后来被人们称作“简单性原则”。正如爱因斯坦所说：“从希腊哲学到现代物理学的整个科学史中，不断有人力图把表面上极为复杂的自然现象归结为几个简单的基本概念和关系。这就是整个自然哲学的基本原理。”这个原理也贯穿了《数学之美》本身。

WWW 的发明人蒂姆·伯纳斯·李谈到设计原理时说过：“简单性和模块化是软件工程的基石；分布式和容错性是互联网的生命。” 虽然在软件工程和互联网领域的从业人员数量极其庞大，但能够真正体会到这些核心思想的人能有多少呢？

我给学生出过这样的考题：把过去十年来重要 IT 杂志的封面上重点推荐的技术专题找来看看，瞧一瞧哪些技术成功了，哪些技术是昙花一现，分析一下原因？其答案很有意思：“有正确设计思想方法的技术”未必能够成功，因为还有非技术的因素；但“没有正确设计思想方法的技术”一定失败，无一例外。因此，我也建议本书的读者结合阅读，体会凝练创造《数学之美》的方法论。

3. 超越欣赏

数学既是对自然界事实的总结和归纳，如英国的哲学家培根所说“一切多依赖于我们把眼睛紧盯在自然界的事实之上”；又是抽象思考的结果，如法国哲学家笛卡尔所说“我思故我在”。这两个方法成就了目前绚丽多彩、魅力非凡的数学，非常值得欣赏。《数学之美》把数学在 IT 领域，特别是语音识别和搜索引擎方面的美丽之处予以了精彩表达。但在这里我想说的是：欣赏美不是终极目的，更值得追求的是创造美的境界。希望本书的读者，特别是年轻读者能够欣赏数学在 IT 技术中的美，学习大师们的思想方法，使自己成为大师，创造新的数学之美。

李星

2012 年 4 月于北京

序言 2

去年我曾经给吴军的《浪潮之巅》写序，今年很高兴得知他的《数学之美》也即将出版了！

和《浪潮之巅》一样，《数学之美》也是当年作为 Google 资深研究员的吴军在谷歌黑板报上应邀撰写的一系列文章。说实在的，刚开始，黑板报的版主还有点担心这个系列会不会让读者觉得太理论而感到枯燥，但很快这个顾虑就被打消了。《数学之美》用生动形象的语言，结合数学发展的历史和实际的案例，谈古论今，系统地阐述了与现代科技领域相关的重要的数学理论的起源、发展及其作用，深入浅出，受到广大读者尤其是科技界人士的喜爱。

之前就曾说过，在我认识的顶尖研究员和工程师里，吴军是极少数具有强大叙事能力和对科技、信息领域的发展变化有很深的纵向洞察力，并能有效归纳总结的人之一。在《数学之美》里，吴军再次展示了这一特点。与《浪潮之巅》不同的是，这次吴军集中阐述了他对数学和信息处理这些专业学科的理解，尤其是他在语音识别、自然语言处理和信息搜索领域多年来的积累 。 从数字和信息的由来，到搜索引擎对信息进行处理背后的数学原理，到与搜索相关的众多领域后面

的奇妙的数学应用，吴军都娓娓道来。他把数学后面的本质思维写得透彻、生动。不得不说，他的文字，引人入胜，也确实让我们体会到数学的美。在他的笔下，数学不是我们一般联想到的枯燥深奥的符号，而是实实在在源于生活的有趣的现象和延伸。数学，其实无处不在，而且有一种让人惊叹的韵律和美！

伽利略曾经说过，“数学是上帝描写自然的语言”；爱因斯坦也曾说过，“纯数学使我们能够发现概念和联系这些概念的规律，这些概念和规律给了我们理解自然现象的钥匙。”我多年来一直也对信息处理、语音识别领域有着一定的研究，深深体会到数学在所有科学领域起到的基础和根本的作用。“哪里有数，哪里就有美”。在这里，我把《数学之美》真诚推荐给每一位对自然、科学、生活有兴趣有热情的朋友，不管你是搞理科还是搞文科的，读一读数学的东西，会让你受益良多，同时能感受到宇宙和世界的美好与奇妙。

吴军把之前谷歌黑板报上的“数学之美”系列文章编辑成现在的这本书，花费了大量的心血和时间。他本着十分严谨的态度，在繁忙的工作之余，补充了之前的系列，并几乎重写了所有的文章，既照顾了普通读者的兴趣，又兼顾了专业读者对深度的要求，很让人钦佩。

有时我在想，现在的社会多了一点压力和浮躁，少了一点踏实和对自然科学本质的好奇求知。吴军的这本《数学之美》真的非常好。非常希望吴军今后能写出更多这样深入浅出的好书，它们会是给这个社会和年轻人最好的礼物。

李开复

2012 年 4 月于北京

前言

数学一词在西方源于古希腊语 μάθημα，意思是通过学习获得的知识的意思，因此早期的数学涵盖的范围比我们今天讲的数学要广得多，和人类的生活也更接近些。在古代最重要的知识，除了对世界的认识和了解，就是人之间的互通和交流了，我们把它称为广义上的通信。本书的内容也将从这里开始。

早期的数学远不如今天神秘，它是非常真实的。但是和任何事物一样，数学也在不断地演化，而这个发展过程使得数学变得高深起来。数学演化的过程实际上是将我们生活中遇到的具体物质以及他们运动的规律不断抽象化的过程。经过几千年的抽象化，大家头脑里能想象的数学只剩下数字、符号、公式和定理了。这些东西和我们的生活似乎渐渐疏远了，甚至在表面上毫不相关了。今天，除了初等数学，大家一般对数学尤其是纯粹数学（Pure Mathematics）的用途甚至产生了怀疑。很多大学生毕业后，在大学所学的数学可能一辈子都没有机会应用，几年后就忘得差不多了。因此，很多人也产生了为什么要学习数学的疑问。更加不幸的是，数学专业的毕业生就连就业也颇为困难，在中国和美国都是如此。在很多人眼里，数学家都是陈景润那样带着厚厚的眼镜、行为木讷的人。因此，无论是这些抽象的数字、符号、

公式和定理，还是研究他们的数学家和美也似乎没有联系。

事实上数学的用途远不止人们的想象，甚至可以说在我们生活中是无所不在。且不说那些和我们生活相对联系较少的领域，比如原子能和航天，那里需要用到大量的数学知识。就说我们天天用的产品和技术，背后都有支持它们的数学基础。作为一名工作了 20 多年的科学工作者，我在工作中经常惊叹于数学语言应用于解决实际问题上时的魔力。我也希望把这种神奇讲解给大家听。

从工业社会起，通信占据了人们生活的大量时间。当人类进入电的时代后，通信的扩展不仅拉近了人与人的距离，而且是带动世界经济增长的火车头。今天通信和它相关的产业可能占到我们世界 GDP 很大的一部分。今天城市里的人花时间最多无非是在电视机前，互联网上，电话上（不论是固定电话还是手机），这些都是这样或者那样的通信。甚至原本必须人到现场的很多活动比如购物，也被建立在现代通信基础之上的电子商务逐渐取代。而现代通信，追溯到 100 多年前的莫尔斯电报码和贝尔的电话，再回到今天的电视，手机和互联网，都遵循信息论的规律，而整个信息论的基础就是数学。如果往更远看，我们自然语言和文字的起源背后都受着数学规律的支配。

“信”字作为“通信”一词的 50%，表明了信息处理存储、传输、处理和理解的重要性。我们今天每个人都使用的搜索，以及我们都觉得很神奇的语音识别、机器翻译和自然语言处理也被包括在其中。也许大家不相信，数学是解决这些问题的最好工具。它不仅能够非常清晰地用一些通用的模型来描述这些领域的看似不同的实际问题，而且能给出非常漂亮的解决办法。每当人们应用数学工具解决一个个和信息处理有关的问题时，总会感叹数学之美。虽然人类的语言有成百上千种，但处理它们的数学模型却是相同的或者相似的，这种一致性也是数学之美的表现。在这本书中，我们将介绍一些数学工具，看看我们是如何利用这些工具来处理信息，开发我们每天生活中都使用的产品。

数学常常给人一种深奥和复杂的感觉，但是它的本质常常是很简单而直接的。英国哲学家弗朗西斯·培根在论美德时讲“美德就如同华贵的宝石，在朴素的衬托下最显华丽。”（Virtue is like a rich stone, best plain set.），数学的妙处也恰恰在于一个好的方法，常常是最简单明了的方法。因此，我会将简单即是美的思想贯穿全书。

最后，要说明一下本书为什么花了相当的篇幅介绍很多我所熟知的自然语言处理和通信的世界级专家。他们来自世界不同的国家，属于不同的民族，但是他们都有一个共同的特点就是数学非常好，同时解决了很多实际问题。通过介绍他们日常的工作和生活，希望读者对真正的世界级学者有更多的了解。了解他们凡人的一面，了解他们成功的原因，了解真正懂得数学之美的人的美好人生。

吴军

2012 年 4 月于深圳

目 录

i **出版说明**

v **序言 1**

ix **序言 2**

xi **前言**

1 **第 1 章　文字和语言 vs 数字和信息**

文字和语言与数学，从产生起原本就有相通性，虽然它们的发展一度分道扬镳，但是最终还是能走到一起。

1　信息
2　文字和数字
3　文字和语言背后的数学
4　小结

15 **第 2 章　自然语言处理 —— 从规则到统计**

人类对机器理解自然语言的认识走了一条大弯路。早期的研究集中采用基于规则的方法，虽然解决了一些简单的问题，但是无法从根本上将自然语言理解实用化。直到 20 多年后，人们开始尝试用基于统计的方法进行自然语言处理，才有了突破性进展和实用的产品。

1　机器智能
2　从规则到统计
3　小结

27 **第 3 章 统计语言模型**

统计语言模型是自然语言处理的基础，并且被广泛应用于机器翻译、语音识别、印刷体或手写体识别、拼写纠错、汉字输入和文献查询。

1 用数学的方法描述语言规律
2 延伸阅读：统计语言模型的工程诀窍
3 小结

41 **第 4 章 谈谈中文分词**

中文分词是中文信息处理的基础，它同样走过了一段弯路，目前依靠统计语言模型已经基本解决了这个问题。

1 中文分词方法的演变
2 延伸阅读：工程上的细节问题
3 小结

49 **第 5 章 隐含马尔可夫模型**

隐含马尔可夫模型最初应用于通信领域，继而推广到语音和语言处理中，成为连接自然语言处理和通信的桥梁。同时，隐含马尔可夫模型也是机器学习的主要工具之一。

1 通信模型
2 隐含马尔可夫模型
3 延伸阅读：隐含马尔可夫模型的训练
4 小结